Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Denis Kuznedelev

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Accurate and efficient methods for neural network compression

Supervisor: Professor Dmitry Yarotsky

Name of the Reviewer: Rebekka Burkholz

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Date: 15-11-2024**

The purpose of this report is to obtain an independent review from the members of PhD defense Jury before the thesis defense. The members of PhD defense Jury are asked to submit signed copy of the report at least 30 days prior the thesis defense. The Reviewers are asked to bring a copy of the completed report to the thesis defense and to discuss the contents of each report with each other before the thesis defense.

If the reviewers have any queries about the thesis which they wish to raise in advance, please contact the Chair of the Jury.

Reviewer's Report

Reviewers report should contain the following items:

- Brief evaluation of the thesis quality and overall structure of the dissertation.
- The relevance of the topic of dissertation work to its actual content
- The relevance of the methods used in the dissertation
- The scientific significance of the results obtained and their compliance with the international level and current state of the art
- The relevance of the obtained results to applications (if applicable)
- The quality of publications
- The summary of issues to be addressed before/during the thesis defense

Thesis Quality and Structure
The dissertation presents a well-organized, coherent, and thorough study on neural network compression, focusing on developing practical methods to improve model efficiency through pruning and quantization. Each chapter builds on previous sections to form a comprehensive exploration of the topic. The dissertation accessible, yet, rigorous. The layout is clear, with each method and result logically following from the previous, contributing to a cohesive narrative which covers complementary approaches to model compression (i.e., pruning, knowledge distillation, and quantization), which are founded in basic optimization theory. Figure 2-2 on Page 43 provides a helpful overview of the thesis structure and how the different approaches come into effect.

Relevance of Topic to Content
The topic of neural network compression is highly relevant to current machine learning and AI challenges, particularly with the surge in deployment of resource-intensive models. This thesis's focus on efficient and accessible model compression addresses a critical area in the field, and each chapter aligns closely with this overarching theme. From introducing methods for unstructured pruning in vision models to fine-tuning techniques for language models, every section is integral to advancing compression strategies applicable across model architectures.

Relevance of Methods Used
The methods employed in this thesis are both theoretically sound and practical. The developed and used techniques such as correlation-aware pruning, Fisher approximation, bilevel quantization, and second-order optimization define state-of-the-art approaches to compress neural networks. Importantly, the thesis applies these methods across various model types, ensuring that the compression techniques are generalizable. This choice of methods aligns well with the goals of efficient and accurate compression, demonstrating relevance both in terms of theoretical rigor and practical utility.

Scientific Significance and International Compliance
The thesis makes significant scientific contributions, particularly in combining established theoretical frameworks, like Optimal Brain Surgeon and Iterative Hard Thresholding, with novel compression methods applicable to modern, large-scale neural networks. The results push the boundaries of neural network compression, providing near-lossless quantization for LLMs and effective pruning for vision models. These contributions align with and advance the international state of the art, as evidenced by the high-quality, peer-reviewed publications in top-tier conferences and journals.

Relevance to Applications
The thesis's results have clear practical applications, especially in resource-constrained environments like mobile devices, autonomous systems, and consumer hardware. Techniques like the sparse-quantized representation (SpQR) and additive quantization for LLMs are particularly relevant to real-world deployment, where model size and computational efficiency are critical constraints. By ensuring that compression does not significantly degrade model accuracy, these methods enhance the feasibility of deploying high-performance neural networks on less capable hardware.

Quality of Publications
The thesis's publication record is impressive, as it is based on seven papers in great machine learning venues, including the A* ranked conferences NeurIPS, ICLR, and ICML. On two of these papers, Denis Kuznedelev is first author. The publications demonstrate both the novelty and the impact of the contributions, reflecting well on his ability to contribute meaningfully to the field.

Summary of Issues to Address Before/During the Thesis Defense
The thesis is comprehensive and does not require any major changes before the defense. Minor issues can be resolved during the defense if necessary.

Provisional Recommendation

☒ *I recommend that the candidate should defend the thesis by means of a formal thesis defense*

☐ *I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate's thesis according to the recommendations of the present report*

☐ *The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense*